DOCUMENT RESUME

ED 115 110 FL 007 280

AUTHOR Freedman, Elaine S.
TITLE Experimentation into Foreign Language Teaching Methodology.
PUB DATE 75
NOTE 27p.; Revised version of a paper given at the Annual Meeting of the British Association for Applied Linguistics (September 1975)

EDRS PRICE MF-$0.76 HC-$1.95 Plus Postage
DESCRIPTORS Academic Achievement; Applied Linguistics; *Audiolingual Methods; *Educational Experiments; French; Grammar; *Language Instruction; Language Laboratories; Pattern Drills (Language); Pronouns; *Second Language Learning; Student Attitudes; *Teaching Methods; Verbs
IDENTIFIERS England

ABSTRACT

This is a preliminary report on a series of small-scale language teaching experiments, aimed primarily at demonstrating that valid research into language teaching methods is possible. Small-scale refers not to the number of subjects involved, but to the scope of the experiment. Instead of looking at a method as a whole (as happens in large-scale global experiments) one limits the area to be investigated, isolating particular variables for study and controlling likely confounding variables. To assess the various methods, two different French grammar topics were presented through the common medium of the language laboratory, at varying levels from first-year university to first-year secondary school in southeast England. Several tapes were made for each topic, all covering the same information, but dealing with it in different ways. The children were divided into groups at random, each child in a particular group using one of the tapes. One group acted as controls. All the subjects were tested on the particular topic before being given the tape, then immediately afterwards, and again ten days later. The results were analyzed statistically, using a computer, to see whether the tapes had had a differential effect on the pupils' achievement and/or attitude scores. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED115110

BRITISH ASSOCIATION FOR APPLIED LINGUISTICS

YORK, SEPTEMBER 1975

EXPERIMENTATION INTO FOREIGN LANGUAGE TEACHING METHODOLOGY

Elaine S. Freedman

FL007280

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Elaine S. Freedman
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

(Revised version of a paper given at the Annual Meeting of the British Association for Applied Linguistics, September 1975.)

Introduction

Large-scale experiments in language teaching, such as that by Scherer and Wertheimer[1] at Colorado between 1960 and 1962, and the Pennsylvania Project[2] (1965-1969) have been widely criticized.

I believe that it is not possible to design a large-scale experiment to improve upon the Pennsylvania Project, although the Peter Green study at York,[3] whilst remaining a 'global' study (in my terms) seems to have achieved a better degree of control, partly through being 'small-scale' in terms of only using 101 subjects. However I still feel that it is the direction of research that must change. It is not a question of finding ways to control the variables in large-scale experiments, since it is the very 'size' or global-ness of the experiment which precludes rigid control. An experiment in which confounding variables are not controlled is not a valid experiment, and a result from an invalid experiment must always be inconclusive.

Many small-scale experiments have been suggested, and attempted, to replace large-scale global comparison ones. 'Small-scale' here refers not to the number of subjects involved, but to the scope of the experiment. Instead of looking at a language teaching method as a whole, one limits the area to be investigated, isolating particular variables for study and controlling confounding variables - a 'specific' rather than a 'global' experiment.

My aim was therefore to show that a small-scale experiment could be carried out in such a way as to produce a valid result. Even if that result showed little difference between groups, or a difference in an unforeseen direction, the result would still be a valid one, because of the way in which it was reached.

As long ago as 1968, Frank Grittner,[4] with the Pennsylvania project in mind, wrote "perhaps we should ask for a cease fire while we search for a more productive means of investigation." I agree wholeheartedly, and I have tried to find that 'means of investigation'.

This paper is an introduction to what is perhaps, in several ways, an unconventional Ph.D. thesis. One might call it a Ph.D. in Language, in Applied Linguistics, in Psychology, in Educational Research or in Experimental Methodology. Each title would in fact be correct. It is an interdisciplinary piece of work, not only in that it draws on different disciplines, but, more especially, in that I have attempted to use the rigours and the methods of one discipline to improve the techniques of another. This is furthermore a reciprocal relationship, with the aim that each discipline does not only draw upon the other disciplines but is actively augmented by them.

Aims

The aim of the work was to ask specific questions about language teaching methodology, and then to attempt to answer them in a valid fashion, but seeking answers was not the only goal. The way of seeking those answers was equally, if not more, important. A great deal of work has already been carried out into comparing the efficacy of various methods of language teaching, but the results have been inconclusive, either because of the means of investigation, or because the students in the different groups tended, in any case, to achieve similar results after several terms' teaching.

Many of the assertions made about the relative merits of different methods have thus been negative rather than positive, i.e. that there was 'no difference' between methods. There is, however, an important

distinction to be made between a result of 'no difference' gained from an experiment which was not valid and the same result gained from one which was.

The work, therefore, had two strands. The first concerned research method, asking the question: 'Is valid research possible?' The second concerned language teaching methodology itself. Here there were three questions, all dealing with the presentation of points of French grammar. The material was presented through the common medium of the language laboratory at varying levels from first-year university to first-year secondary school.

The questions, in general terms, dealt with:

a) Presentation

Teaching the rules of grammar versus teaching by induction from a contextualized scene.

b) Presentation + Practice

Use or non-use of presentation (which one might also call 'explanation').

c) Practice

Contextualized drills versus unrelated drills.

Let us equate PRESENTATION with EXPLANATION. There were 4 conditions:

Fig. 1

			PRACTICE	
PRESENTATION (EXPLANATION)	1	EXAMPLES + RULES	UNRELATED DRILLS	
	2	EXAMPLES + CONTEXTUALIZED SCENE	CONTEXTUALIZED DRILLS	
	3	NONE - PRACTICE DRILLS ONLY	A	UNRELATED DRILLS
			B	CONTEXTUALIZED DRILLS

In question a) one is comparing conditions 1 and 2.

However, before going on to the other two questions, one must look a little more closely at condition 3. There were two kinds of condition 3:

3A Unrelated drills

3B Contextualized drills

Furthermore, it is not strictly true to say that there is no presentation in this condition. There was a rubric for the tape, a minimal presentation so that the students would know how to do the tape - in practical terms. It should be pointed out though that, in the case of the tapes on French Object Pronouns, the pronouns themselves were written out on the front of all tape scripts. One must give the student some idea of the general framework.

In question b) one is comparing

conditions 1 and 3A

2 and 3B

Question c) compares conditions 3A and 3B.

Before describing the actual experiments there are two general theoretical questions which should be mentioned.

Firstly, does one approach research from a theoretical or from a practical point of view?

From the theoretical point of view one says that one expects to find differences between conditions and then one expects to be able to make some kind of theoretical statement about the ordering of those differences. From the practical viewpoint one says that these are the options which exist in schools and one cannot hope to put them into an order except by doing research.

Carroll[5] (1966) opts for the theoretical approach. He presents the choice between the large-scale experiment which '... is feasible but very expensive and difficult to control,' and the 'more precisely controlled, small-scale experiments to check *hypotheses*.' He goes on to assert that '... if research in foreign language teaching is to be really productive, it must become better attuned to theory, both in psychology and in linguistics.'

I find the practical standpoint more acceptable. As I said in 1969,[6] 'Naturally, it is important to know if one method of teaching is generally superior to another or not, but when treating the question from a practical point of view rather than a theoretical one, the problem changes. Instead of asking whether audiolingualism is a useful concept as a whole, one says simply the following: "I want to teach my students certain grammar points. Do I teach them by lecture or by work in the Laboratory?"'

Peter Green (1972)[7] seems to share the view: 'The question was not therefore, "Is the language laboratory effective?" but "Is the teaching in a given situation more effective if the language laboratory is used?"' He also agrees that if a situation actually exists in schools then it is worth investigating.

Perhaps the difference between the theoretical and the practical approaches can be summed up by saying that theoretical infers 'trying to prove' whilst practical means 'trying to find out'. I was 'trying to find out', setting out with certain expectations, but not being so firmly committed to an hypothesis that the design might be biased, however unconsciously, in its favour.

A quotation from Mats Oskarsson (1972),[8] speaking about his GUME project, throws considerable light on the second question, that of the problem of <u>generalization</u>: 'A word of warning against too far-reaching interpretations of the results may be in order. Since our field of

inquiry was restricted to the acquisition of grammar by adults, the findings cannot be automatically carried over to other areas of language learning.'

I would agree entirely with this but I find what follows in the very next paragraph rather alarming, in that Mr. Oskarsson seems to have ignored his own warning:

'The general conclusion that can be drawn from the present study is that adult students acquire foreign-language grammar better by a cognitive method than by a method built exclusively on habit-forming principles.... Finally it can be concluded that the cognitive approach results in better motivation and more favourable attitudes than the habit-forming approach.'

With this in mind, I should like to point out that I have not found answers that will necessarily be valid in any situation other than my own. However, it does not follow that because I would wish to disown Oskarsson's kind of generalization that I believe that there can be no generalization at all. The construction of my experimental sample itself implies a certain degree of generalization.

The first phase of experiments drew 199 students from 6 different schools, and the second phase drew 301 from 7 schools, two of the schools being common to both samples. Unless one accepts a certain amount of generalization, one could never amass a sufficient number of subjects.

I believe that I have thrown some light on the questions I asked. They may not necessarily hold good in other circumstances, but I think that they are <u>interesting</u> answers because they are VALID in my own situation. Although I am not trying to generalize from what I have done, I <u>have</u> found evidence to support generalizations made by other people in the past, e.g. the idea that students' attitudes are linked to the teacher's assessment of their capabilities.

The Experiments

The overall plan of the experiments took the following form. Several language laboratory tapes were made, all dealing with the same grammatical topic, but in different ways. At each school in the experimental sample, the children were divided at random into groups, and each child in a particular group did one of the tapes. One group of children acted as a control and did no tape, although they did do the tests. Each child was tested on the grammar topic before he did the tape, then immediately after it, and again about ten days later. The results were analysed statistically, using a computer, to see whether the tapes had had differential effects on the pupils' achievement and/or attitude scores.

The original plan of the experiment allowed for three different tapes, teaching certain uses of the French subjunctive.

Fig. 2

PRESENTATION	VERBAL EXPLANATION + EXAMPLES	VERBAL EXPLANATION + EXAMPLES	CONTEXTUALIZED SCENE + EXAMPLES
PRACTICE	NON-MEANING-ORIENTED DRILLS	MEANING-ORIENTED DRILLS	MEANING-ORIENTED DRILLS

Before the first pilot run of the experiment the design was changed to 8 tapes. Series 1 and Series 2 drills were constructed to be parallel.

Fig. 3

TAPE	PRESENTATION	+	PRACTICE
I	GRAMMATICAL RULES		UNRELATED DRILLS 1
II	GRAMMATICAL RULES		CONTEXTUALIZED DRILLS 1
III			UNRELATED DRILLS 1
IV			CONTEXTUALIZED DRILLS 1
V	CONTEXTUALIZED SCENE		UNRELATED DRILLS 2
VI	CONTEXTUALIZED SCENE		CONTEXTUALIZED DRILLS 2
VII			UNRELATED DRILLS 2
VIII			CONTEXTUALIZED DRILLS 2

Definitions of these terms might help:

Grammatical rules - a simple explicit statement of the rules governing particular uses of the subjunctive (or, in the second phase, the object pronouns).

Contextualized scene - a naturalistic dialogue which uses the grammar topic a certain number of times in various different ways, so that the students can infer correct usage from these examples.

Contextualized drills - drills built around a situation in such a way that each successive item builds upon the situation. When the series of contextualized drills follows the scene, the original situation for the scene is further developed throughout the drills.

Unrelated drills - have no linking situations between items, or between exercises.

These were the tapes used in the Reading Pilot study in December 1971. Here, as in all further stages, there were three tests for each student taking part.

a) PRE-TEST - attitudes + achievement

b) POST-TEST - attitudes + achievement done immediately after completion of the tape

c) FINAL POST-TEST - achievement done about 10 days after the tape

The subjects in the Reading Pilot were 56 first-year university students, divided amongst the 8 tapes. The results were disappointing. There seemed to be very little difference between the achievement scores on each of the tapes. There were two possible answers. Either there was no difference between the tapes, or the initial scores were so near the ceiling that there was not sufficient room for any improvement at all, let alone differential improvement. If the latter were the answer, then

the materials would have to be made more difficult. As this was not really practicable, the obvious solution was to find students at a lower level.

A shortened pre-test was therefore given to the 6th-form boys of Cranbrook School, and the 6th-form girls of Bournemouth High. These results were encouraging. The average pre-test score in the Reading Pilot was 79.48%. In the Schools Pilot (for the Lower Sixth) this figure came down to 65.81%.

The Schools Pilot test items were then subjected to detailed Item Analysis - rejecting items which were of more than 70% difficulty and less than 0.35 discrimination index (Difference/N/3)[9]. The remaining items were then analysed again, in a fairly subjective manner, to ensure that the different items were evenly reflected in the abridged tests. The remaining contextualized items were then transformed from the remains of 6 different exercises into 6 exercises with one common theme. These were piloted, again in Bournemouth.

The main Subjunctive phase of the experiment was run in six schools between September 1972 and February 1973, using mostly 5th-formers. The 8-tape design turned out to be somewhat impractical, because one needs a large number of students to provide each condition with sufficient subjects. In fact, with 8 tapes there must be 9 conditions, to allow for a control group. Thus in half the schools, only four tapes were used, numbers I, III, VI and VIII, which in fact contained all the major combinations. Once the rejected scripts had been discounted, 199 subjects remained in the sample.

It was thus a '4-tape plus control' design that was chosen for the second phase of experiments. The new topic was the 'French Object

Pronouns', and the target age group was secondary school children at the end of their first year, and about to meet the topic for the first time.

The tapes were now:

Fig. 4

TAPE	PRESENTATION	+	PRACTICE
I	GRAMMATICAL RULES		UNRELATED DRILLS
II			UNRELATED DRILLS
III	CONTEXTUALIZED SCENE		CONTEXTUALIZED DRILLS
IV			CONTEXTUALIZED DRILLS

The tapes, like those in the previous phase, were all recorded using native speakers of French. The test items were piloted in April and May 1973 in Liverpool and in Cranbrook, Kent, on a total of 75 children. The main body of experiments on this topic were run in June and July 1973 in seven schools, ranging from a very academically oriented grammar school to a very rough comprehensive in East London. At one school the experiment was carried out on fourth-year students as a revision exercise, to see whether this would make a difference to the relative usefulness of the tapes. There were thus 361 + 40 students in the Pronouns samples.

Again, the final number of students kept in the sample was very much smaller than the sample originally taken. Extremely ruthless pruning was carried out to make sure that one, the sample was not biased, and two, that there were three completed test scripts for each pupil.

As I have said elsewhere (Freedman, 1971[10]), 'In reality a series of small-scale experiments will prove just as expensive and as difficult to co-ordinate as the one-off large-scale one', and as Dick Allwright stated (1972[11]) 'It must be emphasized that "smaller-scale" does not necessarily imply that fewer subjects need be used for example.' I used 237 subjects for the pilot studies and a total of 500 for the main series of experiments.

In all, this comes to 737 subjects and this figure does not include the 26 teachers who also completed questionnaires. By comparison, the mammoth Pennsylvania Project used 1090 subjects in its second year (Clark, 1969[12]).

Control Measures

One watchword, I believe, of good experimental design should be ELEGANCE, neatness of design, trying to throw light on as many questions as possible by using as few experimental cells or conditions as possible. The other one is CONTROL. Control measures formed a very important part of the experiment, since it was they that made sure that the conditions were identical for each group, with the exception of the tape given.

Firstly, the samples for both series of experiments were drawn from different types of school (comprehensive, grammar, etc.). The tapes were distributed randomly within each school. Matching the groups would have been an impossible task administratively, and random sampling is acknowledged to be extremely efficient.

As for pruning the sample, there were some children who had to be excluded from the final sample because they had obviously not grasped the idea of the tape, for reasons not connected with the tape itself. Some had English language problems that prevented them from coping with the French at all. Some had extremely low IQs. Some children were unpleasant and uncooperative, although the majority responded very well.

The subjunctive experiments dealt with the use of the subjunctive, and so a list of the 40 verbs used in the materials was sent to each school in advance to permit coaching on the relevant forms of the present subjunctive before the experiment. On the scripts themselves, the pupils were told that when in doubt about the precise form, they could just write 'subjunctive'.

The element of control was very important in the writing of the materials. These only used vocabulary found in Français Fondamental 1er degré. In exceptional circumstances 2^{e} degré words or cognates were used. The exercises had to meet strict mathematical constraints regarding the proportions of particular kinds of items, and this constraint combined with that imposed by the contextualization story-line provided considerable problems.

The unrelated and the contextualized test items had to be designed to be parallel, so that the only difference between them would be whether they were unrelated or contextualized. If this had been successfully done, then the pre-test results should have been similar. In the Reading Pilot study, the average score for the contextualized items was 74.34%: that for the unrelated items was 74.6%. Analysis of the exercises in the series also revealed their good internal reliability.

The item analysis of all the items, already mentioned, was another control measure, as was the briefing visit carried out at all schools before each experiment was run.

Using the tapes ensured standardized teaching rather than highly individual teaching, such as occurred in the Pennsylvania Project. The same researcher (myself) was also present at every experimental session.

It was always a question of erring on the side of caution and on the side of the design. It was sometimes hard to make the decisions, but the greater weight was always given to making the experiment valid and reliable. Nonetheless, one cannot overbalance completely on that side, or one's actual questions stop being valid in terms of the situation one is investigating. The chief problem was keeping strict scientific control whilst at the same time not distorting what one was trying to study, and not interfering with sound educational practice.

Data Analysis

Analysis of the data has been carried out on a PDP-10 computer using the language Fortran and SPSS, the Statistical Package for the Social Sciences, first published in 1970 by Nie, Bent and Hull.[13]

The variables in the experiments were divided into background variables, and scores. The background variables included variables like age, sex, attitude to the language laboratory, attitude to interesting content, teacher's attitude to the language laboratory, teacher's assessment of pupil. Together with the different tapes and schools, these comprised the independent variables of the experiment. Measures of attitudes and achievement before and after the tape-lesson comprised the dependent variables - or scores.

The data analysis falls basically into two categories: (A) Descriptive and (B) Inferential.

(A) Descriptive

CODEBOOK provides a description of the values for each variable, both in table and in histogram form. Histograms have the added advantage of permitting before-and-after comparison of the distribution of values, e.g. attitude to the language laboratory before and after exposure to the tape.

CROSSTABS was first used simply for a straightforward crosstabulation of all the variables with all the rest of the variables, to see what there was in the data. Again, this was the 'trying to find out' rather than the 'trying to prove' approach. Later on, CROSSTABS was used in a rather more complex fashion. Individual background variables were each crosstabulated with TAPE, to see the way the values were distributed across the different tapes. This was very useful in providing checks

of the sample, showing that the distribution of the values of the background variables before any experimental intervention did not vary with tape. The next step was to crosstabulate attitudes before with those after the tape, controlling for tape, to see whether the changes in attitudes differed with the tape the subject had been given.

Tests for correlation are usually counted as being a part of inferential statistics, but here they were used more for descriptive purposes. All the correlations were carried out twice over, once using the parametric Pearson Correlation Coefficient test, and again, using the non-parametric Spearman Rank Correlation Coefficient.

First of all, all the subtests of the achievement tests were correlated with each other and with the whole score for the test of which they formed a part. This was done to check the internal reliability of the tests. The various background variables were intercorrelated and they were also correlated with achievement scores.

The highest probability value counted as being statistically significant was $p = 0.05$ or 5%, i.e. there were 5 chances in 100 that the results were due merely to chance. A value of $p = 0.001$ or 0.1% means that there was 1 chance in 1000 that the results were due to chance.

In the ESFPT (pronoun) data, there was a highly statistically significant correlation ($p = 0.001$) between the teachers' assessment of the pupils, and the pupils' own general attitude to languages and French. There was the same degree of correlation ($p = 0.001$) between the pupils' improvement score (post-test minus pre-test score) and their attitude to the language laboratory.

Before going on to the inferential statistics done by computer, it might be worthwhile to report one very interesting phenomenon.

In one grammar school, the sample was drawn from two parallel classes. One was taught by a man greatly in favour of the language laboratory, and the other by a man who was firmly opposed to it. For each pupil an attitude score was calculated from relevant questionnaire items. A Mann-Whitney, non-parametric test for difference revealed a difference significant at $p = 0.001$. From the actual scores it can be seen that the pupils taught by the man who was against the laboratory, were against the tape they had had, irrespective of which tape. Those taught by the pro-laboratory teacher were in favour of their tape, no matter which tape they had had.

It seems that the teacher's attitude had more effect on the pupils' attitudes to the tapes than the tapes themselves.

(B) Inferential Analysis

The programmes BREAKDOWN, ANOVA and ONEWAY all produce similar treatments. ANOVA is a basic analysis of variance and ONEWAY has the advantage of also allowing one to make specific contrasts of different groups within the data.

Before presenting these results perhaps one should define some terms: There are 3 files of data:

Fig. 5

ESFST	Subjunctive series of experiments
ESFPT	Pronoun series of experiments
ESFPTS	Pronoun experiment where the material was used for revision on older children

There are 6 main measures of achievement: PTACWHOL, POTACWHL, FPOTACWH, IMPROVE, OVERALL, FADING.

Fig. 6

PTACWHOL	PRE-TEST WHOLE SCORE
POTACWHL	POST-TEST WHOLE SCORE
FPOTACWH	FINAL POST-TEST WHOLE SCORE
IMPROVE	POST-TEST MINUS PRE-TEST SCORE
OVERALL	FINAL POST-TEST MINUS PRE-TEST SCORE
FADING	FINAL POST-TEST MINUS POST-TEST SCORE

Fig. 7 Breakdown by Tape

	ESFST	ESFPT	ESFPTS
PTACWHOL	Not sig	Not sig	Not sig
POTACWHL	Sig beyond 0.001	Sig beyond 0.001	Not sig
FPOTACWH	Sig at 0.01	Sig beyond 0.001	Not sig
IMPROVE	Sig at 0.001	Sig beyond 0.001	Not sig
OVERALL	Sig at 0.01	Sig beyond 0.001	Not sig
FADING	Not sig	Sig beyond 0.001	Not sig

Fig. 8 Breakdown by School

	ESFST	ESFPT
PTACWHOL	Sig beyond 0.001	Sig beyond 0.001
POTACWHL	Sig beyond 0.001	Sig beyond 0.001
FPOTACWH	Sig beyond 0.001	Sig beyond 0.001
IMPROVE	Not sig	Sig beyond 0.001
OVERALL	Sig at 0.01	Sig beyond 0.001
FADING	Sig at 0.01	Sig at 0.05

Looking at the breakdown by tape, there are no significant differences for any of the measures for the file ESFPTS. This was the pronoun experiment where the material was used for revision only, on 4th-year children, and this result is not surprising. One would not expect the tapes to have much differential effect when the subjects were already quite well-acquainted with the topic. The average pretest score (across all groups) was 60.07%.

There were no significant differences between tapes for the pre-test measures for both the subjunctive (ESFST) and the pronoun (ESFPT) files. This is very encouraging because there really should not be any difference between the groups before the tapes were given. With the exception of the measure 'FADING' (final post-test minus post-test score) for the subjunctive experiments, the differences between groups for the remaining five measures were all statistically significant, most being extremely so.

If one looks at the breakdown by school, all the measures for both files were significant with the exception of the ESFST IMPROVE measure (post-test minus pre-test score).

All the significant differences were then studied in more detail by making contrasts of specific groups within the data.

For the ESFST file, the only measure which produced a significant difference was IMPROVE. The difference between tape I (grammatical rules + unrelated drills) and tape VI (contextualized scene + contextualized drills) was significant at $p = 0.045$, tape I producing higher scores than tape VI.

Fig. 9 ESFPT - Tape Contrasts

MEASURE	SIGNIFICANCE	DIRECTION
POTACWHL	I + III sig beyond 0.001 I + II sig at 0.008	I better than III I better than II
FPOTACWH	I + III sig at 0.001 I + II sig at 0.002	I better than III I better than II
IMPROVE	I + III sig beyond 0.001 I + II sig at 0.005 II + IV sig at 0.043	I better than III I better than II II better than IV
OVERALL	I + III sig beyond 0.001 I + II sig at 0.001	I better than III I better than II
FADING	II + IV sig beyond 0.001	IV better than II

In the ESFPT file, for the four measures, post-test, final post-test, improve and overall, tape I (grammatical rules + unrelated drills) gave significantly better results than tape III (contextualized scene + contextualized drills). This takes us back to question a), that of presentation.

In a climate where contextualized drills are becoming increasingly favourably regarded, it is certainly interesting to see that straightforward grammatical rules appear to produce better immediate results. This takes one back to another of the original questions. In the same way that one can approach the research from either a theoretical or a practical viewpoint, one has the same choice over the interpretation of the results.

The theoretical interpretation would say that the results show that grammatical rules and unrelated drills were superior to contextualized scene and drills. The practical one would say that IN THE EXISTING SITUATION in those schools tape I was superior to tape III. This does not necessarily mean that in the long run, the contextualized material might not turn out to be more effective. In a school where traditional methods are well-established it is not unreasonable to assume that the students

will react more favourably to the kind of material with which they are familiar. Changing that climate might well occur if one carried out more prolonged experimentation in a particular school so that the students were equally well acquainted with both approaches. More of that at the end of the paper.

For these same four measures (post-test, final post-test, improve, overall) too, tape I (grammatical rules + unrelated drills) was significantly superior to tape II (unrelated drills only). This refers to question b), presentation and practice. Not too surprisingly, presentation plus practice produced better results than mere practice alone.

The measure FADING reflects the deterioration of performance over time after the tape. Ebbinghaus' (1885) curve of forgetting[14] shows that learned material is forgotten rapidly at first, and between 6 and 31 days later the amount retained settles from 25% to 21%. The final post-tests were carried out about 10 days after the tape-lesson. It is extremely interesting to see (original question c)) that the scores of those who did tape II (unrelated drills only) decreased, but the scores of those who did tape IV (contextualized drills only) actually improved a little. Perhaps the contextualized material facilitated the longer-term remembering of the grammatical points contained in it.

One vital question remained. Perfectly legitimately, one might ask how one could be sure that these differences were due to the different tapes and not to the various background variables, like students' attitudes, teachers' attitudes, previous use of the laboratory, age, sex, etc. It was for this reason that stepwise MULTIPLE REGRESSION ANALYSIS was carried out on all three files. Multiple regression allows one to study the linear relationship between a set of independent variables and a number of

dependent variables while taking into account the interrelationships among the independent variables. The idea of the steps in the regression is to add progressively the independent variables that one takes into account when trying to explain the dependent variable or score.

From a regression one can predict, say, that for a particular score, e.g. IMPROVE, the basic increase will be 22.67 points. But, if the person went to a technical school, his score would be 22.67 + 26.74. If he did tape I, his score would be 22.67 + 20.29, and if he both went to a technical school and did tape I, one could predict that his score would be 22.67 + 26.74 + 20.29.

The regressions also provided weightings for the various background variables. New standardized measures were then calculated which standardized these background variables within three groups - tapes, schools, and tapes + schools.

One can then say that, for example in the tapes group, the effects of all the variables except the tapes are smoothed out, so that the variables that help improve the score have their improvement subtracted from the score, and variables that hinder the score have their effect added. Thus, if there are still differences, one can say that it is <u>not</u> due to the background variables, which have been standardized, or in other words, neutralized. One can then take the group standardized within tapes and do an analysis of variance for all the measures across the different tape groups. A significant result found <u>now</u> is EXTREMELY likely to be due to the actual tapes themselves. The same may also be done with schools and also with the other statistical procedures carried out before the regressions were done.

The actual analysis is not quite finished, but the analyses of variances <u>do still</u> give significant differences.

Results and Suggestions

To sum up: almost all the measures in the subjunctive experiments and in the main pronoun experiments showed statistically significant differences between the tape-groups. It appears, moreover, that the tape which comprised grammatical rules and unrelated drills was more effective than the one consisting of a contextualized scene and contextualized drills. Grammatical rules plus unrelated drills were also more effective than were the unrelated drills alone. These differences disappeared when the pronoun material was used for revision purposes on more advanced pupils.

The differences in performance found between different types of school were also significant, and both the tapes and the types of school still produced significant differences when adjustments were made to the data to neutralize the effects of the various background variables.

At a later date, I should like to explore further the implications of the results in terms of language teaching methodology itself. However, in this present paper, I think that the emphasis must remain upon the research method itself, and this is why the work has been presented primarily in the form of an experimental report.

In terms of research method then, what might follow from here? I would make two possible suggestions:

a) a series of specific experiments in a chain, so that each part is rigidly controlled, although the chain could be protracted over time.

Fig. 10

Not:

But:

Thus, instead of general teaching by say two different methods over a term or a year, one could do several specific comparisons, building up a term's teaching/experimenting in controllable units.

b) How could small-scale research be made more widely profitable? I would suggest multi-centred but fully co-ordinated small-scale specific experiments. Experiments concentrated in one locality would not really be open to generalization. However, if the original experiment were replicated in sufficient different geographical or L.E.A. (Local Education Authority) regions to establish the 'reliability' of the results, then one could move on to new areas of pedagogical enquiry.

Say:

I. One person does one experiment (or repeats it too).

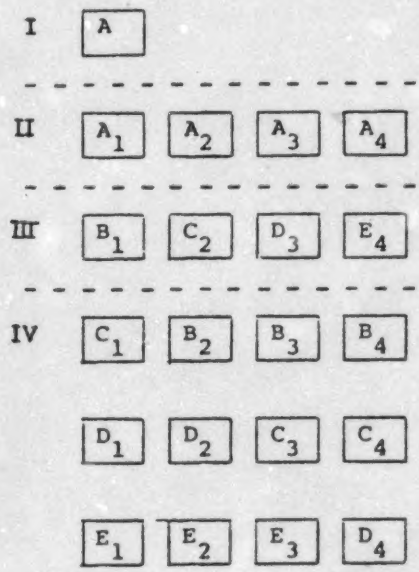
II. Everyone else in the group replicates the experiment.

If the results are reliable then:

III. Various people choose their different options.

IV. Everyone tries those options that look worth pursuing from among the options chosen by the other researchers.

Fig. 11



where A, B, C, D, E = experiments
1, 2, 3, 4 = individual researchers

To come back to my own work. Great problems remain. For example, if one is working with language laboratory drills, the drills that may be good for teaching may not be easily controllable in experiments. I am not saying either that my results are universally applicable to other situations.

How far do experimental situations really reflect the actual teaching situation one finds in schools? Perhaps, having once demonstrated that it is possible to control fully such an experiment, one of the possible next steps could be to begin gradually to relax certain specific items of control, to allow the experiment to mirror more accurately the everyday situation. Knowingly and cautiously relaxing a control measure in a systematic fashion for a specific experimental purpose seems to me far preferable to allowing a haphazard lack of control from the start.

Nonetheless, if the object of valid experimental design is to produce an experiment which gives an answer which can be firmly traced back to the original question, because of the rigidity of control, then I believe that I have achieved such an internally valid design —— a good starting-point.

© Elaine S. Freedman, 1975.
Barnet College and
University of Essex
England

Footnote

It should be pointed out that this paper is a preliminary report, and does not present all the results. There are also many other aspects of the work which are not covered here.

References

1. Scherer, G.A.C. & Wertheimer, M., A Psycholinguistic Experiment in Foreign-Language Teaching (McGraw Hill, 1964).

2. Smith, P.D. Jr., A Comparison of the Cognitive and Audiolingual Approaches to Foreign Language Instruction: The Pennsylvania Foreign Language Project (Center for Curriculum Development, 1970).

3. Green, P.S., "A Comparative study of the effectiveness of the language laboratory in school", in Rôle et Efficacité du Laboratoire de Langues dans L'Enseignement Secondaire et Universitaire, Numéro special 7.3 du Bulletin CILA, 1974, pp. 99-119.

4. Grittner, F.M., "Letter to the Editor", in Newsletter of the National Association of Language Laboratory Directors, Vol. III, No. 2, Dec. 1968, p. 7 (cited by Frank Otto in Modern Language Journal, Vol. 53, No. 6, 1969, p. 420).

5. Carroll, J.B., "The Contributions of Psychological Theory and Educational Research to the Teaching of Foreign Languages", in Trends in Language Teaching, ed. A. Valdman (McGraw Hill, 1966).

6. Freedman, E.S., "An Investigation into the Efficacy of the Language Laboratory in Foreign-Language Teaching", in Audio-Visual Language Journal, Vol. VII, No. 2, 1969, pp. 75-95.

7. Green, P.S., "A Comparative Study of the Effectiveness of the Language Laboratory in School", in AILA Proceedings Copenhagen 1972, Vol. III, (Julius Groos Verlag, 1974), pp. 315-335.

8. Oskarsson, M., "The Acquisition of Foreign Language Grammar by Adults - A Summary Report on Three Field Experiments", in AILA Proceedings Copenhagen 1972, Vol. III, (Julius Groos Verlag, 1974), pp. 520-535.

9. Ingram, E., "Appendix : Item Analysis", in Language Testing Symposium, ed. A. Davies (Oxford University Press, 1968), pp. 192-203.

10. Freedman, E.S., "The Road from Pennsylvania - where next in language teaching experimentation?", in Audio-Visual Language Journal, Vol. IX, No. 1, 1971, pp. 33-38.

11. Allwright, R.L., "Prescription and Description in the Training of Language Teachers", in AILA Proceedings Copenhagen 1972, Vol. III, (Julius Groos Verlag, 1974), pp. 150-166.

12. Clark, J.L.D., "The Pennsylvania Project and the 'Audio-Lingual versus Traditional' Question", in Modern Language Journal, Vol. 53, No. 6, 1969, p. 389.

13. Nie, N., Bent, D.H. & Hull, C.H., Statistical Package for the Social Sciences , (McGraw Hill, 1970).

14. Ebbinghaus, H., see Experimental Psychology by Woodworth, R.S. & Schlosberg, H. (Methuen, 1954), p. 726.